به نام خدا

تمرین اول درس داده‌کاوی

ترم بهار ۱۴۰۲

توضیحات تمرین:

- پاسخ به این تمرین به صورت انفرادی می‌باشد.
- لطفا سوالات را به ترتیب پاسخ دهید.
- در صورت ابهام درباره‌ی تمرین با ایمیل درس با تدریسیاران در ارتباط باشید.

dm.spring1402@gmail.com

- همچنین می‌توانید سوالات خودتان را از طریق تلگرام و آیدی‌های زیر مطرح کنید.

@armanhtm

@iamelmo

- مهلت ارسال تمرین تا **ساعت ۱۱:۵۹ دقیقه روز جمعه ۱۸ فروردین** می‌باشد.
- تمرین شامل دو بخش **تئوری و عملی** می‌باشد.
- فایل‌های ارسالی شما باید یک فایل pdf گزارش (شامل جواب سوالات تئوری و سوالات بخش عملی)، و همچنین شامل کدهای شما باشد، که لطفا آنها را تحت یک فایل zip بارگذاری نمایید.
- فرمت فایل zip شما باید به شکل زیر باشد:

HW1-[student_number].zip (برای مثال HW1-9800000.zip)

فهرست

بخش تئوری

سوال اول

به سوالات زیر پاسخ دهید.

الف) داده‌ی پرت[1] با نویز[2] را با یکدیگر مقایسه کنید.

ب) یک سناریو بیان کنید که در آن داده‌های پرت برای ما مفید هستند و اطلاعات ارزشمندی از آن دریافت می‌کنیم.

ج) مشخص کنید که آیا یک نویز می‌تواند داده‌ی پرت باشد یا خیر؟

سوال دوم

در حوزه‌ی داده‌کاوی، انبار داده[3] چیست و چه تفاوت و شباهتی با پایگاه داده[4] دارد؟

سوال سوم

یکی از روش‌های یافتن داده‌های پرت استفاده از توزیع نرمال[5] و percentile ها است. در مورد این روش تحقیق کرده و آن را توضیح دهید.

سوال چهارم

فرایند پاکسازی داده‌ها[6] و نمایش داده‌ها[7] را در نظر بگیرید:

الف) فرایند پاکسازی داده‌ها را تعریف کنید.

ب) اهمیت نمایش داده‌ها را بیان کنید و به یک مورد از چالش‌های آن اشاره کنید.

ج) چرا پاکسازی داده‌ها یک فرایند مهم و پیشنیاز برای نمایش داده‌ها می‌باشد؟

---

[1] outlier
[2] noise
[3] data warehouse
[4] database
[5] Normal distribution
[6] data cleaning/cleansing
[7] data visualization

سوال پنجم

در یک آزمایشگاه ژنتیک مقدار فعالیت دو ژنوم مختلف مورد بررسی قرار گرفته و در ۱۰ بازه زمانی مختلف در به صورت زیر ثبت شده است:

Gen\time	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
G1	-3	5	8	-2	1	2	3	-5	10	-1
G2	9	20	16	8	2	10	-6	-15	25	-2

الف) با استفاده از معیار شباهت Cosine Similarity، Correlation، Mutual Information شباهت این دو ژن را مقایسه کنید.

ب) طبق نتایج هر معیار مشخص کنید آیا دو ژنوم از یکدیگر مستقل هستند یا خیر.

ج) آیا نتایج به دست آمده متفاوت است؟ اگر پاسخ مثبت است علت آن را توضیح دهید.

سوال ششم

دو مورد از روش‌های data preprocessing روش‌های aggregation و sampling هستند. این دو روش را توضیح داده و مزایا و معایب هر یک را بنویسید.

سوال هفتم

در رابطه با کاهش بعد تحقیق کرده و به سوالات زیر پاسخ بدهید.

الف) مفاهیم انتخاب ویژگی[8]، استخراج ویژگی[9] و مهندسی ویژگی[10] را توضیح و تفاوت‌های بین آن‌ها را بیان کنید.

ب) الگوریتم‌های کاهش بعد به دو دسته خطی و غیرخطی تقسیم می‌شوند. تفاوت این دو دسته را توضیح داده و روش کار الگوریتم PCA از دسته خطی و الگوریتم t-sne از دسته غیرخطی را توضیح دهید.

[8] Feature selection
[9] Feature extraction
[10] Feature engineering

سوال هشتم

برای داده‌های عددی زیر نمودار جعبه[11] را رسم کنید.

۴۱, ۲۸, ۳۹, ۲۵, ۴۹, ۳۶, ۲۷, ۳۳, ۲۶, ۷۰, ۲۷, ۲۹, ۱, ۳, ۲۷

سوال نهم

همانطور که می‌دانید، یکی از روش‌های مقایسه دو توزیع آماری استفاده از روش q-q plot است.

الف) نحوه کار این روش را توضیح دهید.

ب) نمودار q-q plot می‌تواند به شکل‌های متفاوتی نمایان شود: به طور مثال شبیه یک خط راست مورب. سه نوع از این شکل‌های متفاوت را بررسی کنید و تحلیل خود داده‌های توزیع‌های آماری ورودی به آن را بنویسید. به نظر شما از روی شکل q-q plot چه مواردی در مورد توزیع‌های آماری اولیه قابل استنتاج است؟

سوال دهم

برای هر یک از روش‌های نرمال‌سازی زیر تحقیق کرده و بازه‌ی اعداد را مشخص کنید.

الف) نرمال‌سازی min-max

ب) نرمال‌سازی z-score

ج) نرمال‌سازی با مقیاس‌دهی[12]

سوال یازدهم

با توجه به مقادیر ورودی X و مقادیر هدف Y می‌توان یک برازش خطی یا غیرخطی بر روی بسیاری از دادگان‌ها ایجاد کرد. با توجه به این مقادیر، به سوالات زیر پاسخ دهید.

$$X = [2, 4, 1, 3, 2, 6], \qquad Y = [5, 6, 3, 6, 3, 10]$$

الف) روش محاسبه معادله نرمال[13] را با استفاده از روش محاسبه مشتق جزئی باقی‌مانده[14] کامل شرح دهید.

ب) یک برازش خطی ($Y = \beta_1 X + \beta_0$) را برای این دادگان محاسبه کنید. (مقدار خطای برازش را نیز به دست آورید)

ج) یک برازش غیرخطی ($Y = \beta_2 X^2 + \beta_1 X + \beta_0$) برای این دادگان محاسبه کنید. (مقدار خطای برازش را نیز به دست آورید)

[11] Box plot
[12] decimal scaling
[13] Normal Equation
[14] residual

بخش پیاده‌سازی

پیش‌پردازش داده‌ها برای مدل‌های یادگیری ماشین یک مهارت اصلی برای هر مهندس یادگیری ماشین و همینطور هر دانشمند داده است. به طور کلی دو کتابخانه مطرح pandas و scikit-learn برای پیش‌پردازش داده‌ها استفاده می‌شود که ما در این بخش به بررسی این کتابخانه‌ها می‌پردازیم.

در یک پروژه علم داده در دنیای واقعی، پیش‌پردازش داده‌ها یکی از مهمترین گام‌های آن است و یکی از عوامل مشترک موفقیت یک مدل می‌باشد. یعنی اگر پیش‌پردازش داده‌ها و مهندسی ویژگی‌ها به درستی انجام پذیرد، احتمال موفقیت آن مدل در مقایسه با مدلی که داده‌ها برای آن به خوبی پیش‌پردازش نشده اند، بیشتر است و نتایج بهتری تولید خواهد کرد.

مجموعه داده[15]

مجموعه داده در نظر گرفته شده برای این تمرین Palmer penguin می‌باشد. این مجموعه برای شناسایی سه نژاد مختلف پنگوئن (Adelie، Gentoo و Chinstrap) جمع‌آوری شده است. برای هر پنگوئن ۷ ویژگی وجود دارد که در ادامه توضیحات ویژگی‌ها قرار داده شده اند.

- island: island name (Dream, Torgersen, or Biscoe) in the Palmer Archipelago (Antarctica)
- bill _length_mm: bill length (mm)
- bill_depth_mm: bill depth (mm)
- flipper_length_mm: flipper length (mm)
- body_mass_g: body mass (g)
- sex: penguin sex
- species: penguin species

مجموعه داده تحت یک فایل csv در اختیار شما قرار داده شده است. لازم است در ابتدا آن را بارگذاری نمایید.

[15] dataset

داده‌های از دست رفته

یک عبارت معروف در یادگیری ماشین وجود دارد که ممکن است آن را شنیده باشید:

Garbage in, Garbage out.

اگر مجموعه داده شما مملو از مقادیر NaN باشد، مدل نیز نتیجه‌ی قابل قبولی ندارد. بنابراین مقابله با چنین داده‌هایی مهم است.

سوال اول – ابتدا به دنبال داده‌های NaN در مجموعه داده بگردید و ذکر کنید که از هر ویژگی چند سطر فاقد داده هستند. برای این کار از تابع ()isna استفاده کنید.

اگر تعداد سطرهای با مقادیر از دست رفته کم باشد، یا داده‌های ما به گونه‌ای باشند به گونه‌ای که توصیه نمی‌شود مقادیر از دست رفته را پر کنیم، می‌توانم آن سطر را حذف کنیم (برای مثال داده‌هایی مرتبط با سلامت افراد را در نظر بگیرید. اگر میزان فشار خون یک فرد مشخص نشده بود و جای آن خالی بود، ما نمی‌توانیم آن را با میانگین فشار خون بقیه افراد پر کنیم).

سوال دوم – داده‌های از دست رفته در مجموعه داده را با استفاده از ()dropna حذف کنید. و تعداد سطرهای مجموعه داده را قبل و بعد از حذف عنوان کنید.

حال می‌خواهیم به جای حذف، داده‌های از دست رفته را پر کنیم. یک روش برای پر کردن مقادیر از دست رفته، پر کردن آن با میانگین، میانه و یا واریانس آن ستون است. برای انجام این کار می‌توانیم از SimpleImputer از sklearn استفاده کنیم. البته در این موارد باید داده‌های ما عددی باشند. برای داده‌های غیر عددی یکی از ساده‌ترین راه‌ها پر کردن آن مقدار با متداول‌ترین مقدار آن ستون می‌باشد.

سوال سوم – در گام اول داده‌های عددی از دست رفته در مجموعه داده را با میانگین آن ستون جایگزین کنید (یعنی تنها برای ویژگی‌های: bill _length_mm، bill_depth_mm، flipper_length_mm، body_mass_g). در گام دوم داده‌های غیرعددی از دست رفته (island، sex، species) را با متداول‌ترین مقدار جایگزین کنید.

داده‌های غیرعددی

به طور کلی در علم داده، مدل‌های ما قادر به درک یک داده‌ی متن نیستند و لازم است که این داده‌ها به عدد تبدیل شوند. برای تبدیل ویژگی‌های کلاس‌بندی شده می‌توان از دو روش Label Encoding و یا One Hot Encoding استفاده کرد.

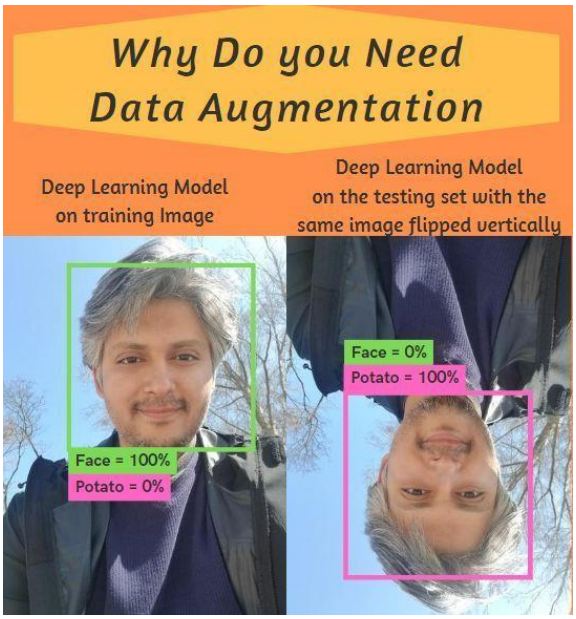
سوال چهارم – با استفاده از Label Encoding در ستون‌های زیر، تغییرات را اعمال کنید.

- در ستون Island: Biscoe را به ۰، Dream را به ۱، Torgersen را به ۲ تبدیل کنید.
- در ستون sex: female را به ۰ و male را به ۱ تبدیل کنید.
- و در ستون species: Adelie را به ۰، Chinstrap را به ۱ و Gentoo را به ۲ تبدیل کنید.

می‌توانید این کار را با کمک sklearn.preprocessing انجام دهید.

افزایش داده‌ها[16]

این تکنیک برای چندین هدف متفاوت به کار می‌رود. به طور مثال به تصویر زیر نگاه کنید.



از سویی دیگر از جمله مشکلاتی که در بسیاری از پروژه‌های مربوط به هوش مصنوعی و علم داده وجود دارد، محدود بودن دیتاست و یا نامتوازن بودن تعداد داده‌ها در هر کلاس است. این مشکلات سبب ایجاد اختلال در عملکرد شبکه می‌شوند. یکی از روش های حل این مشکل استفاده از data augmentation است .

سوال پنجم - نحوه عملکرد این روش چگونه است و تبدیل هایی که در آن استفاده می‌شود را شرح دهید. آیا از این روش برای داده‌های تست استفاده می‌شود؟ علت را شرح دهید.

سوال ششم - روش‌های upsampling و downsampling و ترکیبی را توضیح دهید.

سوال هفتم - در مورد روش‌های smotetomek و smoteenn تحقیق کنید نحوه کار آن‌ها را توضیح دهید. وجه اشتراک این دو روش چیست؟

سوال هشتم - از داده‌ها آموزش یکی از کلاس‌های دیتاست داده شده ۹۰ درصد را حذف کنید. حال با استفاده از این دو روش غیر متعادل بودن دیتاست که با حذف کردن داده‌ها بوجود آوردیم را هندل کنید. (نیازی به پیاده‌سازی این دو روش نیست، و میتوانید از کتابخانه‌ها استفاده کنید).

این سوال تنها برای آموزش افزایش داده‌ها بوده و در گام‌های بعدی با دیتاست اصلی کار کنید.

[16] data augmentation

نرمال‌سازی

از آزمایش‌های مشخصی ثابت شده است که مدل‌های یادگیری ماشین و یادگیری عمیق در مقایسه با مجموعه داده‌هایی که نرمال-سازی نشده‌اند، در یک مجموعه داده نرمال، عملکرد بهتری دارند. هدف از نرمال‌سازی تغییر مقادیر به یک مقیاس مشترک است. چندین راه برای این کار وجود دارد.

سوال نهم – با استفاده از StandardScaler در sklearn.preprocessing اقدام به نرمال‌سازی داده‌ها کنید. مقدار واریانس و میانگین هر ستون را قبل و بعد از نرمال‌سازی ذکر کنید (دقت کنید که این نرمال‌سازی را بر روی برچسب‌ها (ستون species) انجام ندهید).

تحلیل مولفه‌های اصلی[17]

برای بسیاری از پروژه‌های یادگیری ماشین، تجسم داده‌ها به درک بهتر پروژه کمک می‌کند. تجسم داده‌های ۲ یا ۳ بعدی چندان چالش برانگیز نیست. همچنین در بعضی از پروژه‌های یادگیری ماشین، ویژگی‌های استخراج شده، ویژگی‌های اضافی هسند و می‌توان آنها را کاهش داد. تحلیل مولفه‌های اصلی یا همان PCA به ما کمک می‌کند تا بردار ویژگی‌های خود را از یک فضای n بعدی به k بعدی تبدیل کنیم.

سوال دهم – با استفاده از PCA در sklearn.decomposition مولفه‌های اصلی داده‌ها را حساب کنید و بردار ویژگی‌ها از یک فضای ۶ بعدی به ۳ بعدی کاهش دهید (پیش‌نیاز این کار، نرمال‌سازی داده‌ها است).

مصورسازی[18]

همانطور که در قسمت قبل گفته شد، تجسم داده‌ها برای فهم بهتر پروژه به ما کمک خواهد کرد. در این قسمت اقدام به رسم داده‌های مجوعه داده خود خواهیم کرد. برای رسم ویژگی‌ها از سه ویژگیِ استخراج شده در قسمت قبل استفاده کنید.

سوال یازدهم – با استفاده از کتابخانه‌ی matplotlib داده‌های مجموعه داده را رسم کنید. دقت کنید برای ویژگی‌ها از ویژگی‌های حاصل از PCA استفاده کنید (رسم شکل به صورت سه‌بعدی خواهد شد). برای هر کلاس رنگ متفاوتی در نظر بگیرید.

سوال دوازدهم – برای هر ۶ ویژگی ارائه شده در مجموعه داده، نمودار boxplot را رسم کنید (این کار را قبل از گامِ نرمال‌سازی انجام دهید).

موفق و پیروز باشید.

تیم تدریسیاری درس داده‌کاوی – بهار ۱۴۰۲

[17] principal component analysis
[18] visualization